AgentStealth: Reinforcing Small Language Models for On-Device Text Anonymization

Anonymous ACL submission

Abstract

Casual user-generated content often harbors linguistic cues that adversaries may exploit to infer sensitive attributes, posing significant privacy risks. Existing anonymization methods either rely on simple keyword replacement that degrades content utility or leverage cloud-based LLMs, which introduce risks of secondary privacy leakage. Locally deployed Small Language Models (SLMs) offer a promising alternative, but balancing privacy protection and data utility remains difficult due to the scarcity of high-quality supervision. To address this, we propose AgentStealth, a self-reinforcing framework for robust text sanitization with preserved semantic integrity. First, an adversarial multi-agent system enhanced by *In-context Contrastive Learning* and *Adaptive Utility-Aware Control* generates high-quality anonymization data. Second, we adapt SLMs using this data, incorporating both anonymization and attack signals to foster dual-role proficiency. Finally, online reinforcement learning uses internal adversarial feedback to iteratively strengthen the model's defense against profiling attacks. Experiments demonstrate that AgentStealth significantly outperforms baselines in anonymization success rate (+27.7 percentage points) and data utility preservation (+0.37). By enabling lightweight deployment, AgentStealth facilitates on-device text anonymization for privacy-sensitive applications. Our code is open-source at `https://anonymous.4open.science/r/AgentStealth-E168`.

1 Introduction

In today's digital landscape, social media and online platforms have empowered users globally to communicate and share in real time. With nothing more than a smartphone or computer, individuals can freely post comments and content from virtually anywhere. While such user-generated texts often appear casual or spontaneous, they frequently embed latent cues that may inadvertently disclose sensitive personal attributes, such as age, gender, geographic location, or marital status (European Union, 2016). These implicit signals can be exploited by adversaries to infer private information (Brewster, 2023; Wu et al., 2025). For instance, prior work (Staab et al., 2024) demonstrates that even zero-shot LLMs can accurately predict user attributes from text, presenting a virtually barrier-free privacy attack surface. Recent autonomous agents like OpenClaw (Steinberger, 2026), which have full access to personal data, are even more susceptible. Such risks pose a serious threat to individual privacy and highlight the urgent need for effective text anonymization methods (McCallister et al., 2010).

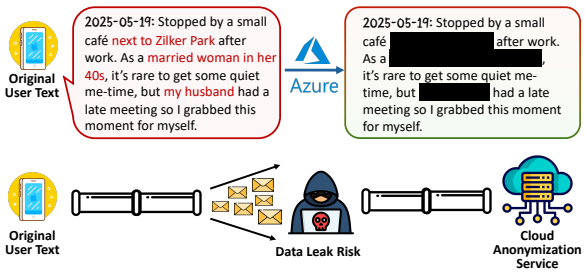


Figure 1: Limitations of existing anonymization methods: rule-based substitution severely degrades usability, and cloud-based methods introduce additional risks.

To address this issue, various anonymization tools have been developed (Presidio, 2025), such as Azure (Aahill, 2023). These tools typically rely on rigid entity recognition combined with rule-based substitution strategies. Although easy to implement, they often result in overly aggressive redactions that severely undermine the text's interpretability and communicative value, failing to preserve the necessary *data utility* for meaningful online interaction, as illustrated in Figure 1. Here, "data utility" refers to the extent to which the anonymized text preserves the original meaning, readability, and fluency. Recent research has explored using LLMs as anonymization agents (Frikha et al., 2024), demonstrating

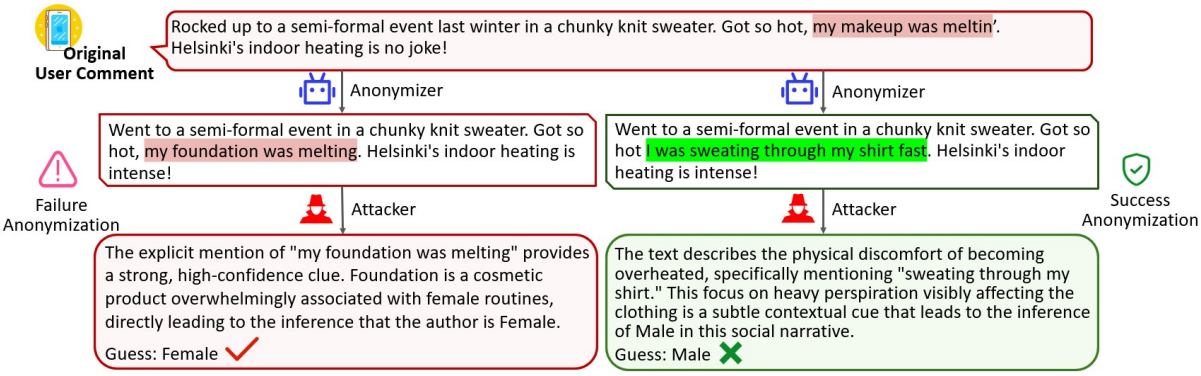


Figure 2: One successful case and one failed case of text anonymization using LLMs.

superior data utility compared to conventional methods (Staab et al., 2025). Figure 2 illustrates this paradigm by presenting both a successful anonymization instance and a failure case. Nevertheless, current approaches predominantly rely on large-scale models hosted in the cloud, introducing three critical drawbacks that hinder responsible deployment. First, uploading user data to the cloud introduces secondary privacy risks, as the security of data transmission and storage cannot always be guaranteed. Second, centralized cloud inference incurs substantial computational costs and latency (Gim et al., 2025), limiting accessibility. Third, existing metrics often overlook whether the anonymized text supports users' practical needs for authentic expression, which is vital for maintaining fairness in online communities.

To mitigate these risks, we propose strictly on-device text anonymization using locally deployed Small Language Models (SLMs). This completely avoids external transmission vulnerabilities. However, off-the-shelf SLMs struggle with anonymization, and optimizing them is impeded by the prohibitive cost and scarcity of high-quality training data. To address this challenge, we propose **AgentStealth**, a self-reinforcing anonymization framework that leverages a comprehensive three-stage training pipeline to empower SLMs with robust privacy capabilities.

First, we establish an adversarial anonymization multi-agent system enhanced by *In-context Contrastive Learning* to distill insights from contrasting success and failure cases (Zhao et al., 2024), and by *Adaptive Utility-Aware Control* to safeguard data utility during anonymization. Second, we perform supervised fine-tuning (SFT) on the SLM using high-quality data synthesized by this system, incorporating both anonymization and attack signals. This joint training cultivates the model's dual proficiency as both a privacy defender and an attribute attacker. Finally, an online reinforcement learning (RL) stage further refines the SLM's performance. By leveraging real-time adversarial feedback from its own fine-tuned attack capabilities, the model iteratively improves its defense against self-identified vulnerabilities.

We conduct experiments on two representative datasets (Yukhymenko et al., 2024; Staab et al., 2024). Results show that our trained SLM achieves state-of-the-art anonymization performance across two backbone models, outperforming baseline methods with an average anonymization success rate increase of up to 27.7 percentage points and an average absolute score gain of up to 0.37 in data utility preservation. Moreover, the model's lightweight architecture enables efficient deployment on edge devices. By confining sensitive data processing to the local devices, we eliminate dependencies on cloud servers, effectively closing the attack surface exposed during data transmission. Our key contributions are summarized as follows:

- We propose **AgentStealth**, a self-reinforcing LLM anonymization framework that iteratively enhances privacy protection through a novel pipeline combining high-quality data synthesis, dual-role SFT and self-adversarial RL.
- We introduce an innovative multi-agent system incorporating *In-context Contrastive Learning* to extract actionable insights from historical outcomes, and *Adaptive Utility-Aware Control* to dynamically balance anonymization with utility.
- Through extensive evaluation, we validate that our lightweight SLM can effectively replace cloud-based LLMs for anonymization tasks, enabling a decentralized, privacy-preserving solution suitable for on-device deployment.

2 Related Works

Privacy Risks with the Use of LLMs: With the rapid application of LLMs across various domains,

privacy concerns have increasingly emerged (Yao et al., 2024; Shao et al., 2025; Li et al., 2024; Wang et al., 2025b; Yuan et al., 2024; Cai et al., 2024). Attackers can extract training data (Kim et al., 2023), determine if personal data was used for fine-tuning (Kandpal et al., 2024), or even compromise the memory of LLM-based agents (Wang et al., 2025a) through crafted prompts. Additionally, cloud-hosted LLMs expose user queries directly to providers, prompting research into privacy-preserving cloud invocation methods (Zhang et al., 2024a,b,c). Finally, even without using cloud services, malicious actors can leverage LLMs to infer Personally Identifiable Information (PII) implicitly contained in users' social media comments (Staab et al., 2024; Frikha et al., 2024; Staab et al., 2025).

LLM-powered Author Profiling: Risks and Defenses Author profiling, aiming to infer personal attributes from texts, is a long-standing NLP task traditionally relying on machine learning to classify limited attributes like age and gender (Pardo et al., 2018; Rosso et al., 2016). Recently, LLMs have been shown to effectively infer a much broader range of attributes (e.g., occupation, location) through textual analysis (Staab et al., 2024). To address the resulting privacy leakage, LLM-based Adversarial Anonymization methods have been developed and validated to protect texts against such attacks while partially preserving data utility (Staab et al., 2025; Frikha et al., 2024).

RL for Enhancing LLM Reasoning LLM inference can be naturally formulated as a RL problem, where the model learns a policy to maximize cumulative rewards over token sequences (Xu et al., 2025). Recently, RL has emerged as a powerful approach to improve LLM reasoning (Shao et al., 2024) by utilizing outcome-driven reward functions that align reasoning trajectories with final objectives, rather than merely predicting the next token. Furthermore, methods like GRPO (Guo et al., 2025) improve training stability and efficiency by assigning relative rewards across sampled outputs, avoiding the need for a separate critic model.

3 Problem Formulation

In this section, we formally define the task of anonymizing textual data to protect sensitive attributes while preserving data utility. Let $t \in \mathcal{T}$ denote an input text containing private information, and $a \in \mathcal{A}$ represent its corresponding sensitive attribute (e.g., location, age, or gender). The anonymization model M_{anony} transforms the original text t into an anonymized form $\tilde{t} = M_{\text{anony}}(t)$, while the attacker model M_{attack} attempts to infer the sensitive attribute $\tilde{a} = M_{\text{attack}}(\tilde{t})$.

The attack accuracy over the dataset $\mathcal{T}$ is defined as the proportion of correctly inferred attributes, given by $Acc_{\text{attack}} = \frac{1}{|\mathcal{T}|}\sum_{(t,a)\in\mathcal{T}} \mathbb{I}[M_{\text{attack}}(M_{\text{anony}}(t)) = a]$, where $\mathbb{I}[\cdot]$ is the indicator function. Our goal is to minimize Acc_{attack} such that the attacker cannot reliably recover a from $\tilde{t}$.

Simultaneously, the anonymized text must maintain high data utility $\mathcal{U}$ with respect to the original content. We measure this through a set of semantic similarity metrics:

$$\mathcal{U} = \frac{1}{N}\sum_{i=1}^{N} \text{sim}_i(t, M_{\text{anony}}(t)), \quad (1)$$

where $\text{sim}_i(\cdot)$ includes BLEU, ROUGE, and other similarity scores. To jointly optimize for privacy protection and utility preservation, we define a composite objective function $\mathcal{J}$ to balance the trade-off:

$$\mathcal{J} = \lambda \cdot (1 - Acc_{\text{attack}}) + (1 - \lambda) \cdot \mathcal{U}, \quad (2)$$

where $\lambda \in [0, 1]$ is a tunable hyperparameter. A higher λ prioritizes stronger anonymization, while a lower value favors higher data utility retention.

4 Methods

As shown in Figure 3, our approach can be divided into three sequential stages. Firstly, we construct a comprehensive multi-agent system to perform privacy attack and protection, which is enhanced by *In-context Contrastive Learning* and *Adaptive Utility-Aware Control*. This system enables the generation of a large volume of high-quality, task-specific data containing both anonymization and attack signals, which we use for SFT. This joint training enhances the model's dual capabilities as both defender and attacker. Building on this, we apply RL where the model leverages its own real-time adversarial feedback to iteratively improve anonymization performance. During inference, the fine-tuned model is deployed to perform text anonymization tasks.

4.1 Multi-agent Anonymization System with Insight Memory

Inspired by the adversarial anonymization method proposed by Staab et al.(Staab et al., 2025), we design a novel anonymization pipeline to collect high-quality data for training. The input dataset is first divided into a set of batches $\mathcal{B} = \{B_1, B_2, \ldots, B_m\}$,

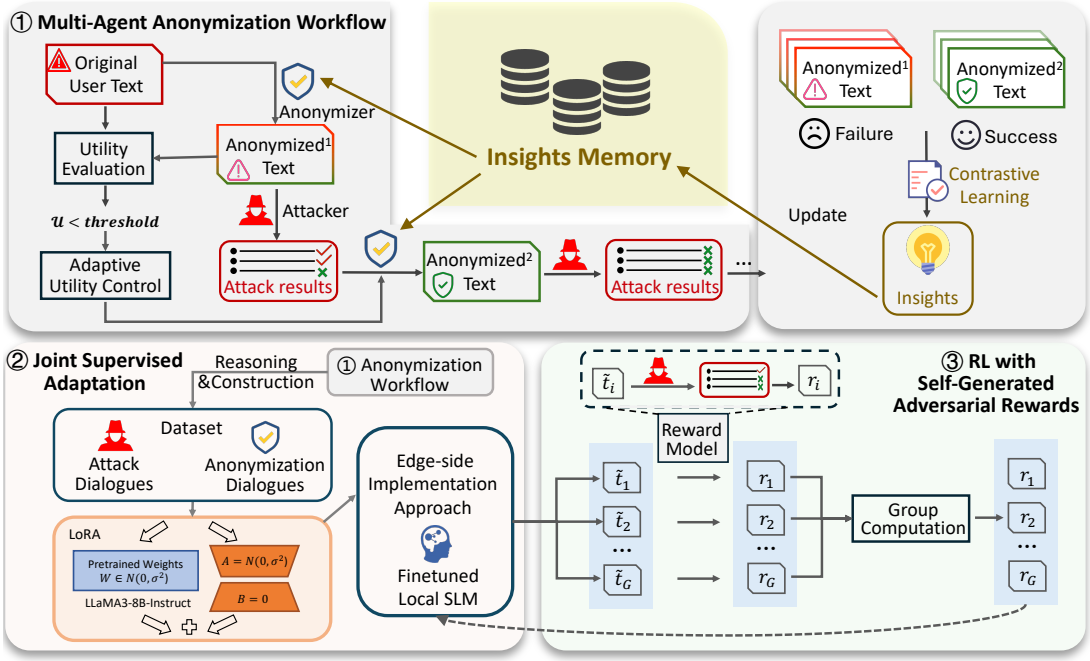


Figure 3: Illustration of AgentStealth Framework.

where each batch $B_j = \{(t_k, a_k)\}_{k=1}^{K_j}$ consists of K_j text-attribute pairs. A memory module $\mathcal{M}_{\text{mem}}$ is initialized as empty and gradually updated during training; it stores anonymization insights I for different types of PII. For each batch B_j, we anonymize the input texts using the adversarial anonymization process, guided by both the current insight memory $\mathcal{M}_{\text{mem}}$ and an adaptive prompting strategy P_{adapt}, which is integrated into the multi-agent system to better preserve data utility. The anonymized output at k-th iteration is denoted as

$$\tilde{t}^{(k)} = M_{\text{anony}}^{(k)}(t \mid \mathcal{M}_{\text{mem}}, P_{\text{adapt}}). \quad (3)$$

After processing each batch, the memory module is updated as $\mathcal{M}_{\text{mem}} \leftarrow \text{Update}(\mathcal{M}_{\text{mem}}, B_j, \text{outcomes}_j)$, leveraging both successful cases and success-failure pairs to refine future anonymization behavior.

4.1.1 In-context Contrastive Learning

There is a well-known saying: "Learn from your mistakes." In the context of text anonymization, both successful and failed anonymization attempts carry valuable insights. Successful examples offer strategies that effectively preserve privacy, while failures serve as cautionary signals, exposing vulnerabilities that should be avoided. Motivated by this, we introduce *In-context Contrastive Learning*, which systematically extracts knowledge from historical outcomes to enhance future anonymization performance. During the data construction phase, we have already divided the dataset into batches and apply the full adversarial anonymization and attack pipeline. Let the protection status of a text t with sensitive attribute a, anonymized as $\tilde{t}$, be defined as: $S(t, a, \tilde{t}) = 1 - \mathbb{I}\left[M_{\text{attack}}(\tilde{t}) = a\right]$, where $\mathbb{I}[\cdot]$ is the indicator function returning 1 if the predicted attribute matches the ground truth and 0 otherwise. A protection success corresponds to $S = 1$, while a failure yields $S = 0$. Then we identify samples that change protection status during the N-step adversarial interaction process. For instance, if a sample transitions from failure to success between steps i and j, we extract a contrastive pair: $(\tilde{t}_{\text{fail}}, \tilde{t}_{\text{success}}) = \left(M_{\text{anony}}^{(i)}(t), M_{\text{anony}}^{(j)}(t)\right)$, where $S(t, a, \tilde{t}_{\text{fail}}) = 0$ and $S(t, a, \tilde{t}_{\text{success}}) = 1$. If no success-failure pairs can be collected within a batch, then only successful examples are retained.

Specifically, LLM is guided by an automatically constructed prompt P_{contrast} to produce an insight $I_{\text{new}} = LLM_{\text{reason}}(P_{\text{contrast}}, \tilde{t}_{\text{fail}}, \tilde{t}_{\text{success}})$, which is a concise and generalizable explanation of why the second anonymization succeeded while the first failed. To ensure the system adapts over time, we maintain a memory module updated via $\mathcal{M}_{\text{mem}} \leftarrow \text{SelectTopK}(\mathcal{M}_{\text{mem}} \cup \{I_{\text{new}}\}, M_{\max})$, where $\text{SelectTopK}(\cdot)$ retains only top-$M_{\max}$ most valuable insights according to predefined criteria.

This rolling update scheme alleviates the cold-start issue and allows continuous refinement of anonymization strategies during training. Notably, since ground truth labels are unavailable during inference, the experience buffer $\mathcal{M}_{\text{mem}}$ is frozen

post-training, and all downstream evaluations rely solely on the stored set of insights. Appendix A.6 provides a case study showing the optimization process of the insights.

4.1.2 Adversarial Anonymization with Adaptive Utility-Aware Control

Previous anonymization methods often overlooked the importance of preserving data utility, leading to outputs that, while privacy-preserving, were poorly suited for downstream tasks. In contrast, our proposed multi-agent system explicitly integrates a utility-aware mechanism that dynamically adjusts the anonymization strategy based on data utility feedback. This enables the model to maintain the data utility and intent of the original text throughout the anonymization process. Specifically, after each round of adversarial anonymization, which produces $\tilde{t}^{(k)}$, we evaluate the data utility of the anonymized text relative to the original input t. The data utility score is computed as $\mathcal{U}(\tilde{t}^{(k)}, t) = \frac{1}{N_{\text{sim}}} \sum_{i=1}^{N_{\text{sim}}} \text{sim}_i(t, \tilde{t}^{(k)})$, where $\text{sim}_i(\cdot)$ includes BLEU (Papineni et al., 2002), ROUGE-1 and ROUGE-L (Lin, 2004).

Then we assess whether the data utility degradation, defined as $(\mathcal{U}(\tilde{t}^{(k-1)}, t) - \mathcal{U}(\tilde{t}^{(k)}, t))$, exceeds a predefined threshold $\tau_{\mathcal{U}}$. If the degradation is too large, a warning is embedded into the next prompt to guide the model toward preserving the original intent and structure more faithfully. The adaptive prompt $P_{\text{adapt}}^{(k+1)}$ for $(k+1)$-th round is defined as:

$$P_{\text{adapt}}^{(k+1)} = \begin{cases} P_{\text{base}} \oplus W_{\mathcal{U}} & \text{if } \Delta\mathcal{U}^{(k)} > \tau_{\mathcal{U}}, \\ P_{\text{base}} & \text{otherwise,} \end{cases} \quad (4)$$

where $\Delta\mathcal{U}^{(k)} = \mathcal{U}(\tilde{t}^{(k-1)}, t) - \mathcal{U}(\tilde{t}^{(k)}, t)$ denotes the utility drop, P_{base} is the standard anonymization prompt, $W_{\mathcal{U}}$ is a data utility warning message, and $\oplus$ denotes prompt concatenation. This adaptive prompting strategy empowers the model to apply more aggressive anonymization when data utility is preserved, while issuing corrective signals only when degradation is detected. By avoiding rigid instructions at every step, the model retains flexibility while ensuring that usability remains intact. The anonymization model then proceeds with the updated prompt and accumulated experience.

4.2 Joint SFT with Anonymization and Adversarial Signals

Our anonymization multi-agent system generates a comprehensive dataset $\mathcal{D}_{\text{SFT}}$ for supervised adaptation of one SLM. This dataset is designed to enhance the SLM's capabilities in two distinct yet complementary roles: as a privacy protector (defender) and as an attribute attacker. $\mathcal{D}_{\text{SFT}}$ comprises: (1) Anonymization data: pairs $(t_i, \tilde{t}_i^*)$, where t_i is an original text and $\tilde{t}_i^*$ is its high-quality anonymized version. This includes instances hardened against previously identified vulnerabilities by incorporating analysis of attack cases, where t_i might be augmented with contextual information from such analyses, and $\tilde{t}_i^*$ is the robustly anonymized output. (2) Attack data: pairs (t'_j, a_j), where t'_j is a text (which could be an original, partially anonymized, or fully anonymized text) and a_j is the sensitive attribute the model learns to infer from t'_j.

By training on this diverse dataset, the SLM, denoted as $M'_{\text{dual}}(\cdot; \theta_{SFT})$ with parameters θ_{SFT}, simultaneously develops proficiency in both generating privacy-preserving text and identifying sensitive attributes from text. This joint training significantly enhances its data utility and security posture, effectively cultivating its dual capabilities as both a defender and an attacker. The objective of SFT is to minimize a cross-entropy loss $\mathcal{L}_{\text{SFT}}(\theta_{SFT}) = \sum_{(x_k, y_k) \in \mathcal{D}_{\text{SFT}}} \text{Loss}(M'_{\text{dual}}(x_k; \theta_{SFT}), y_k)$ across all examples. Here, (x_k, y_k) represents a generic input-output pair from $\mathcal{D}_{\text{SFT}}$, where x_k is the input text (potentially augmented with task-specific instructions or context derived from attack analyses) and y_k is the target output (either a robustly anonymized text $\tilde{t}_i^*$ or a sensitive attribute a_j). This dual-focus SFT prepares the model for the subsequent RL stage where its own SFT-enhanced attack capabilities can be leveraged for further refinement.

4.3 Reinforcement Learning with Self-Generated Adversarial Rewards

Following SFT, we further refine the model's *anonymization* capabilities, now denoted $M'_{\text{anony}}(\cdot; \theta_{RL})$ (with parameters θ_{RL} initialized from θ_{SFT}) through RL. Crucially, the SFT stage has already equipped the model $M'_{\text{dual}}(\cdot; \theta_{SFT})$ with strong attribute inference (attack) abilities. We leverage this inherent capability by using the attack function of $M'_{\text{dual}}(\cdot; \theta_{SFT})$, denoted $M'_{\text{attack}}(\cdot; \theta_{SFT})$, as the source of real-time adversarial feedback during RL. This self-adversarial setup means the model effectively uses its own SFT-enhanced proficiency as an attacker to provide instructive feedback, eliminating the need for a separate, external attacker model and allowing the

anonymizer to improve its defenses against its own refined attack strategies. The primary objective of this RL phase is to bolster the model's privacy protection performance.

We adopt an online RL setup leveraging the Group Reward Policy Optimization (GRPO) algorithm (Shao et al., 2024). The reward signal R is designed to primarily optimize for privacy protection (anonymity), potentially balanced with data utility. It is formulated as a weighted sum:

$$R(t, \tilde{t}, a) = \lambda_{RL} \cdot R_{\text{anonymity}}(\tilde{t}, a) + (1 - \lambda_{RL}) \cdot R_{\text{utility}}(\tilde{t}, t), \tag{5}$$

where $\tilde{t}$ is the anonymized text generated by $M'_{\text{anony}}(\cdot; \theta_{RL})$ from the original text t which has sensitive attribute a, and $\lambda_{RL} \in [0, 1]$ is a tunable hyperparameter. The **AnonymityReward**, $R_{\text{anonymity}}$, quantifies the success of anonymization by reflecting the failure rate of the SFT-enhanced internal attacker: $R_{\text{anonymity}} = 1 - \mathbb{I}[M'_{\text{attack}}(\tilde{t}; \theta_{SFT}) = a]$.

This component directly encourages policies that hinder the model's own advanced attack capabilities. The **UtilityReward** (R_{utility}) measures the preservation of content quality and semantic meaning, $R_{\text{utility}}(\tilde{t}, t) = \mathcal{U}(\tilde{t}, t)$, as defined previously (Equation 1). In our experiments, we set $\lambda_{RL} = 0.5$ by default, thereby balancing anonymity and data utility objectives in the reward function. Reward $\{r_i\}$ for each output o_i is computed using Equation 5. Based on the above defined rewards, the model can be optimized following the standard GRPO procedure. The training details are provided in Appendix A.8.

5 Experiments

We evaluate the proposed LLM-based anonymizer in terms of privacy–utility trade-off, out-of-domain generalization and alignment with human preferences. Additional qualitative case studies are provided in Appendix A.6 and Appendix A.7. All prompts are provided in Appendix A.9.

5.1 Settings

Datasets

We conduct our experiments using two synthetic datasets: (1) the SynthPAI Reddit comment corpus (Yukhymenko et al., 2024); and (2) 525 synthetic Q&A pairs (Staab et al., 2024). Both datasets contain texts annotated with eight personal attributes (age, gender, location, occupation, education, relationship status, income, and birth place), and have been shown to closely mirror authentic user-generated content (Staab et al., 2024, 2025; Yukhymenko et al., 2024). To eliminate privacy risks associated with real data, we exclusively employ these synthetic datasets for training and release. Given their similarity, we merge them, allocating 200 samples (100 from each) for testing and the remainder for training.

Models During the multi-agent system construction phase, we employ the DeepSeek-V3 model (DeepSeek-AI, 2024), while for local model training and deployment, we opt for the Llama-3.1-8B-Instruct model (Meta, 2024) and Qwen-2.5-1.5B-Instruct model (Team, 2024). All models are open-source, facilitating transparency and reproducibility in our experiments.

Hyperparameters For each attribute mentioned in Section 4.1.1, we set the number of most valuable insights to 5 in order to reduce cost. We applied LoRA in both the SFT and RL stages. Details for SFT and RL are provided in Appendix A.5. Detailed sensitivity analyses for the RL reward weight λ and LoRA rank are deferred to Appendix A.4.

5.2 Evaluation

Metrics We establish a dual-aspect assessment protocol that measures both the efficacy of privacy protection and the preservation of data utility. For privacy quantification, we define **Anonymity** as the proportion of attributes where the top-1 predicted entity in the fifth stage anonymized outputs diverges from ground truth, and **Progress** as the proportion of attributes that exhibit enhanced privacy preservation compared to their original unprotected forms. Complementing these security metrics, we assess functional data utility through: (a) semantic similarity metrics: BLEU (Papineni et al., 2002), ROUGE-1, ROUGE-L (Lin, 2004); (b) LLM-based **Readability** scoring (DeepSeek-V3, on a scale from 1-10) evaluating linguistic fluency; and (c) LLM-based **Meaning** scoring (DeepSeek-V3, on a scale from 1-10), quantifying similarity of meaning between original and anonymized texts. To quantitatively measure the data utility levels, we compute the average **Utility Score** using:

$$Score_{\text{Utility}} = \frac{1}{5}\left[S_{\text{B}} + S_{\text{R1}} + S_{\text{RL}} + \frac{S_{\text{read}} - 1}{9} + \frac{S_{\text{mean}} - 1}{9}\right] \tag{6}$$

where S_{B}, S_{R1}, and S_{RL} denote the BLEU, ROUGE-1, and ROUGE-L scores, respectively. S_{read} and S_{mean} represent the Readability and Meaning preservation scores. These metrics simultaneously capture privacy gains and data utility

Methods	Anonymity	Progress	BLEU	ROUGE-1	ROUGE-L	Readability	Meaning	$Score_{\text{Utility}}$
Standard Prompt	17.2%	11.3%	0.27	0.32	0.32	3.98	3.86	0.31
AA	37.8%	23.7%	0.37	0.53	0.52	6.40	5.70	0.51
IncogniText	50.6%	32.6%	0.55	0.69	0.68	7.84	6.65	0.66
AgentStealth	62.2% (+11.6%)	43.1% (+10.5%)	0.60 (+0.05)	0.83 (+0.14)	0.82 (+0.14)	9.68 (+1.84)	8.47 (+1.82)	0.81 (+0.15)
w/o SFT	49.6%	30.9%	0.73	0.88	0.88	9.61	8.79	0.86
w/o RL	60.5%	42.9%	0.60	0.82	0.81	9.55	8.38	0.80
w/o SFT&RL	48.3%	27.9%	0.69	0.86	0.85	9.12	8.31	0.82

Table 1: Main results of AgentStealth and baselines for Llama-3.1-8B-Instruct. Higher is better for all metrics.

Methods	Anonymity	Progress	BLEU	ROUGE-1	ROUGE-L	Readability	Meaning	$Score_{\text{Utility}}$
Standard Prompt	4.6%	3.4%	0.00	0.01	0.01	1.13	1.08	0.01
AA	4.4%	1.9%	0.04	0.06	0.06	1.59	1.54	0.06
IncogniText	4.8%	1.5%	0.06	0.08	0.08	1.88	1.75	0.08
AgentStealth	48.5% (+43.7%)	31.7% (+28.3%)	0.49 (+0.43)	0.68 (+0.60)	0.67 (+0.59)	8.26 (+6.38)	7.06 (+5.31)	0.67 (+0.59)
w/o SFT	4.4%	2.7%	0.03	0.05	0.05	1.50	1.47	0.05
w/o RL	47.0%	32.3%	0.45	0.64	0.63	8.07	6.90	0.64
w/o SFT&RL	4.2%	3.2%	0.03	0.05	0.04	1.52	1.47	0.05

Table 2: Main results of AgentStealth and baselines for Qwen-2.5-1.5B-Instruct. Higher is better for all metrics.

trade-offs in the anonymization process.

Baselines We evaluate AgentStealth against three baselines: (i) Standard Prompt for conventional anonymization, and (ii) the Adversarial Anonymization method (AA) proposed by Staab et al (Staab et al., 2025), and (iii) the IncogniText method proposed by Frikha et al (Frikha et al., 2024). We employ two foundational models: Llama-3.1-8B-Instruct and Qwen-2.5-1.5B-Instruct, evaluating their performance under the three baselines.

5.3 Overall Performance

We present our main results in Table 1 and Table 2. The experimental results demonstrate that **AgentStealth** surpasses all baselines and ablation studies for both SLMs. Compared with the strongest baselines, the Llama-based **AgentStealth** achieves a 11.6 percentage points improvement in **Anonymity** (from 50.6% to 62.2%) while demonstrating an average 0.15 enhancement in $Score_{\text{Utility}}$ (from 0.66 to 0.81). For Qwen, the anonymization performance improves by 43.7 percentage points (from 4.8% to 48.5%) while demonstrating an 0.59 enhancement in $Score_{\text{Utility}}$ (from 0.08 to 0.67). This is primarily because the untuned 1.5B model often fails to consistently adhere to the required structured output formats, consequently leading to unextractable output. A detailed evaluation of second-stage SFT is provided in Appendix A.1. Also the attack performances at different training stages are shown in Appendix A.2.

5.4 Privacy–Utility Trade-off

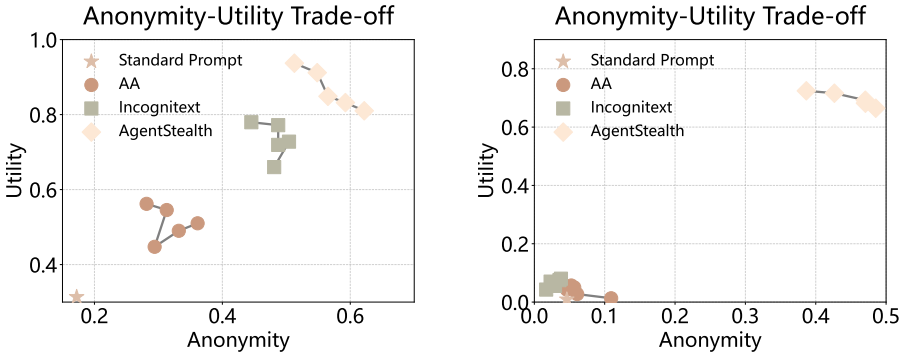
In text anonymization, anonymity and data utility are inherently competing objectives, as stronger anonymization often compromises usability for downstream tasks. To illustrate this trade-off, we plot the metric **Anonymity** against $Score_{\textbf{Utility}}$. We vary the number of adversarial iterations for the baselines (AA and IncogniText) to control this balance. For **AgentStealth**, evaluating the anonymized outputs at each iteration of its adaptive utility-aware prompting forms a trade-off curve, demonstrating its flexibility for real-world privacy requirements.As shown in Figure 4, **AgentStealth** outperforms the baselines across both SLMs, achieving higher data utility at equivalent anonymity levels. Specifically, with Llama-3.1-8B-instruct, our method establishes a dominant frontier. **AgentStealth** operates within a high Anonymity range ($X \approx 0.51$ to 0.62) while maintaining utility scores between $Y \approx 0.81$ and 0.93. In contrast, the best-performing baseline, IncogniText, peaks at a maximum Anonymity of $X \approx 0.50$ with a corresponding utility of $Y \approx 0.72$, proving our approach more effectively optimizes the privacy-utility tradeoff.

5.5 Human Preference Evaluation

To evaluate the data utility of different anonymization methods, we conducted a human evaluation

Dataset	Methods	Anonymity	Progress	BLEU	ROUGE-1	ROUGE-L	Readability	Meaning	Score Utility
Synthetic Dialogue Evaluation Set	Standard Prompt	1.5%	0.0%	0.06	0.10	0.10	1.96	1.96	0.09
	AA	2.3%	0.8%	0.09	0.12	0.11	2.10	2.08	0.11
	Incognitext	0.0%	0.0%	0.0	0.0	0.0	1.0	1.0	0.0
	AgentStealth	22.9% (+20.6%)	23.7% (+22.9%)	0.30 (+0.21)	0.47 (+0.35)	0.46 (+0.35)	5.87 (+3.77)	5.10 (+3.02)	0.45 (+0.34)
Blog Authorship Corpus	Standard Prompt	4.0%	3.5%	0.03	0.04	0.04	1.43	1.37	0.04
	AA	5.3%	5.1%	0.04	0.07	0.07	2.01	1.71	0.08
	IncogniText	7.1%	4.0%	0.05	0.08	0.08	1.95	1.67	0.08
	AgentStealth	25.3% (+18.2%)	23.7% (+18.6%)	0.37 (+0.32)	0.55 (+0.47)	0.52 (+0.44)	7.06 (+5.05)	5.79 (+4.08)	0.53 (+0.45)

Table 3: Evaluation of AgentStealth's generalization capability on unseen datasets. Higher is better for all metrics.



(a) Privacy-Utility Trade-off for Llama-3.1-8B-Instruct

(b) Privacy-Utility Trade-off for Qwen-2.5-1.5B-Instruct

Figure 4: Comparison of Privacy-Utility Trade-off Between Different Models

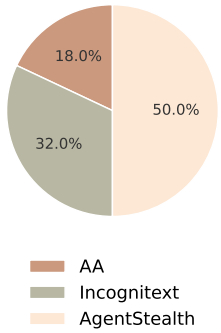


Figure 5: Human Evaluation

where five computer science students provided a total of 150 responses. Participants were presented with identical source texts anonymized by AA, IncogniText, and **AgentStealth** in a random order. They then made pairwise blind comparisons to select the optimal substitution based on meaning similarity. As shown in Figure 5, our method achieves higher human evaluation metrics than both baselines, demonstrating its practical effectiveness.

5.6 Generalization Evaluation

To evaluate **AgentStealth**'s generalization capability, we employed a dual-evaluation strategy. First, we generated 60 novel author profiles using Yukhymenko's data generation approach (Yukhymenko et al., 2024) and extracted one utterance per profile from simulated dialogues to form the 60-sample Synthetic Dialogue Evaluation Set. Second, to test robustness against domain shifts and real-world text diversity, we introduced the Blog Authorship Corpus (Schler et al., 2006). We then tested **AgentStealth** and baseline approaches on these unseen datasets using Qwen-2.5-1.5B-Instruct.

As shown in Table 3, **AgentStealth** consistently outperforms all baselines in anonymization. Compared with the strongest baselines, our method achieves an improvement of 20.6 percentage points (2.3% to 22.9%) on the synthetic set, and a gain of 18.2 percentage points (7.1% to 25.3%) on the Blog Authorship Corpus. These results demonstrate that our fine-tuned method maintains strong generalization on entirely unseen datasets.

See Appendix A.3 for sentiment similarity analysis further validating core semantic preservation.

6 Conclusions

This work enhances text anonymization on edge devices to prevent adversaries from inferring users' personal attributes. We develop an effective multi-agent system integrating two core components: an *In-context Contrastive Learning* mechanism to extract insights from prior outcomes, and an *Adaptive Utility-Aware Control* module to preserve data utility. To address edge device constraints, we construct a high-quality reasoning dataset to fine-tune lightweight, locally deployed SLMs. By operating entirely on-device, our approach keeps sensitive data under user control, eliminating privacy risks associated with cloud communication. Additionally, we leverage reinforcement learning to further boost performance, establishing a novel paradigm for on-device privacy-preserving text processing.

Limitations

While AgentStealth demonstrates strong performance in on-device text anonymization, its evaluation is currently constrained by the scarcity of high-quality, open-source personal attribute datasets due to inherent privacy restrictions. Consequently, our reliance on primarily synthetic data means its robustness and generalizability in highly specialized, real-world domains—such as medical or legal contexts—remain to be fully validated. Furthermore, while our framework effectively distills defense strategies into an insight memory during training, this memory is frozen post-deployment. As a result, the local model currently lacks a mechanism to continuously adapt to novel, emerging profiling attacks in real-time, presenting a valuable direction for future on-device continual learning research.

References

Aahill. 2023. What is azure ai language - azure ai services. Accessed: 2025-05-12.

Thomas Brewster. 2023. Chatgpt has been turned into a social media surveillance assistant.

Kunlin Cai, Jinghuai Zhang, Zhiqing Hong, William Shand, Guang Wang, Desheng Zhang, Jianfeng Chi, and Yuan Tian. 2024. Where have you been? a study of privacy risk for point-of-interest recommendation. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 175–186.

DeepSeek-AI. 2024. Deepseek-v3 technical report. *Preprint*, arXiv:2412.19437.

European Union. 2016. General data protection regulation (gdpr) – legal text. `https://gdpr-info.eu/`. Accessed: 2025-05-15.

Ahmed Frikha, Nassim Walha, Krishna Kanth Nakka, Ricardo Mendes, Xue Jiang, and Xuebing Zhou. 2024. Incognitext: Privacy-enhancing conditional text anonymization via llm-based private attribute randomization. In *Neurips Safe Generative AI Workshop 2024*.

In Gim, Caihua Li, and Lin Zhong. 2025. Confidential prompting: Protecting user prompts from cloud llm providers. *Preprint*, arXiv:2409.19134.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Nikhil Kandpal, Krishna Pillutla, Alina Oprea, Peter Kairouz, Christopher Choquette-Choo, and Zheng Xu. 2024. User inference attacks on large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18238–18265.

Siwon Kim, Sangdoo Yun, Hwaran Lee, Martin Gubri, Sungroh Yoon, and Seong Joon Oh. 2023. Propile: probing privacy leakage in large language models. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pages 20750–20762.

Songwei Li, Jie Feng, Jiawei Chi, Xinyuan Hu, Xiaomeng Zhao, and Fengli Xu. 2024. Limp: Large language model enhanced intent-aware mobility prediction. *arXiv preprint arXiv:2408.12832*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Erika McCallister, Timothy Grance, and Karen Scarfone. 2010. Guide to protecting the confidentiality of personally identifiable information (pii). Special Publication 800-122, National Institute of Standards and Technology, Gaithersburg, MD.

Meta. 2024. Introducing llama 3.1: Our most capable models to date.

Dat Quoc Nguyen, Thanh Vu, and Anh-Tuan Nguyen. 2020. Bertweet: A pre-trained language model for english tweets. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 9–14.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Francisco M. Rangel Pardo and 1 others. 2018. Overview of the 6th author profiling task at PAN 2018: Multimodal gender identification in twitter. In *CLEF 2018 Working Notes*, volume 2125 of *CEUR Workshop Proceedings*. CEUR-WS.org. 6th Author Profiling Task at PAN@CLEF 2018.

Presidio. 2025. Get serious about cybersecurity.

Paolo Rosso, Francisco Rangel Pardo, Martin Potthast, Efstathios Stamatatos, Michael Tschuggnall, and Benno Stein. 2016. Overview of PAN'16. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, volume 9822 of *Lecture Notes in Computer Science*, pages 332–350, Cham. Springer.

Jonathan Schler, Moshe Koppel, Shlomo Argamon, and James W Pennebaker. 2006. Effects of age and gender on blogging. In *AAAI spring symposium: Computational approaches to analyzing weblogs*, volume 6, pages 199–205. Stanford, CA, USA.

Chenyang Shao, Xinyuan Hu, Yutang Lin, and Fengli Xu. 2025. Division-of-thoughts: Harnessing hybrid language model synergy for efficient on-device agents. In *Proceedings of the ACM on Web Conference 2025*, pages 1822–1833.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Robin Staab, Mark Vero, Mislav Balunović, and Martin Vechev. 2024. Beyond memorization: Violating privacy via inference with large language models. In *International Conference on Learning Representations 2024*.

Robin Staab, Mark Vero, Mislav Balunovic, and Martin Vechev. 2025. Language models are advanced anonymizers. In *The Thirteenth International Conference on Learning Representations*.

Peter Steinberger. 2026. OpenClaw: Your own personal AI assistant. Open source autonomous AI agent framework.

Qwen Team. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Bo Wang, Weiyi He, Pengfei He, Shenglai Zeng, Zhen Xiang, Yue Xing, and Jiliang Tang. 2025a. Unveiling privacy risks in llm agent memory. *CoRR*.

Huandong Wang, Wenjie Fu, Yingzhou Tang, Zhilong Chen, Yuxi Huang, Jinghua Piao, Chen Gao, Fengli Xu, Tao Jiang, and Yong Li. 2025b. A survey on responsible llms: Inherent risk, malicious use, and mitigation strategy. *arXiv preprint arXiv:2501.09431*.

Shujin Wu, Yi R Fung, Cheng Qian, Jeonghwan Kim, Dilek Hakkani-Tur, and Heng Ji. 2025. Aligning llms with individual preferences via interaction. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 7648–7662.

Fengli Xu, Qianyue Hao, Zefang Zong, Jingwei Wang, Yunke Zhang, Jingyi Wang, Xiaochong Lan, Jiahui Gong, Tianjian Ouyang, Fanjin Meng, and 1 others. 2025. Towards large reasoning models: A survey of reinforced reasoning with large language models. *arXiv preprint arXiv:2501.09686*.

Yifan Yao, Jinhao Duan, Kaidi Xu, Yuanfang Cai, Zhibo Sun, and Yue Zhang. 2024. A survey on large language model (llm) security and privacy: The good, the bad, and the ugly. *High-Confidence Computing*, page 100211.

Hanyang Yuan, Jiarong Xu, Cong Wang, Ziqi Yang, Chunping Wang, Keting Yin, and Yang Yang. 2024. Unveiling privacy vulnerabilities: Investigating the role of structure in graph data. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 4059–4070.

Hanna Yukhymenko, Robin Staab, Mark Vero, and Martin Vechev. 2024. A synthetic dataset for personal attribute inference. In *Advances in Neural Information Processing Systems*, volume 37, pages 120735–120779. Curran Associates, Inc. NeurIPS 2024.

Kaiyan Zhang, Jianyu Wang, Ermo Hua, Biqing Qi, Ning Ding, and Bowen Zhou. 2024a. Cogenesis: A framework collaborating large and small language models for secure context-aware instruction following. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4295–4312.

Mengke Zhang, Tianxing He, Tianle Wang, Lu Mi, Niloofar Mireshghallah, Binyi Chen, Hao Wang, and Yulia Tsvetkov. 2024b. Latticegen: Hiding generated text in a lattice for privacy-aware large language model generation on cloud. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2674–2690.

Xinyu Zhang, Huiyu Xu, Zhongjie Ba, Zhibo Wang, Yuan Hong, Jian Liu, Zhan Qin, and Kui Ren. 2024c. Privacyasst: Safeguarding user privacy in tool-using large language model agents. *IEEE Transactions on Dependable and Secure Computing*.

Andrew Zhao, Daniel Huang, Quentin Xu, Matthieu Lin, Yong-Jin Liu, and Gao Huang. 2024. Expel: Llm agents are experiential learners. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19632–19642.

A Appendix

A.1 Comprehensive Evaluation of SFT

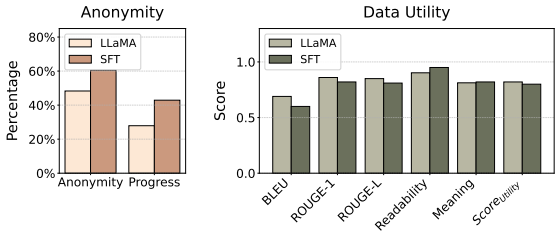


Figure 6: Effectiveness of SFT for Llama-3.1-8B-Instruct

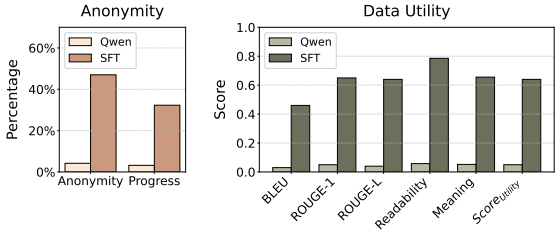


Figure 7: Effectiveness of SFT for Qwen-2.5-1.5B-Instruct

To evaluate the effectiveness of the second-stage SFT, we conducted a comparative analysis between the original models before SFT and the models after SFT. The evaluation was carried out from two dimensions: anonymity and data utility. The experimental results reveal that SFT can significantly improve anonymization performance by 19.2% with Llama and by 34.2% with Qwen, as shown in Figure 6 and Figure 7 (all data utility scores are normalized to the range [0, 1]).

A.2 Evaluation of Attack Performance

We present the enhanced attack performance of the two underlying base models, Llama and Qwen, when augmented by the **AgentStealth** framework, visualizing the results at different training stages in Figure 8. The experimental results clearly demonstrate the significant efficacy and portability of our framework. Specifically, when **AgentStealth** is implemented atop the Llama base model, it achieves a substantial improvement in attack accuracy by 15 percentage points (from 50% to 65%), effectively achieving performance comparable to the strong DeepSeek-V3 model. Even more remarkably, when deployed with the Qwen base model, the framework realizes an even greater performance boost, registering an impressive improvement of 25 percentage points (from 31% to 56%). This consistent and dramatic improvement across different foundational models underscores the robustness of our method in maximizing attack performance, irrespective of the initial capabilities of the underlying SLM.

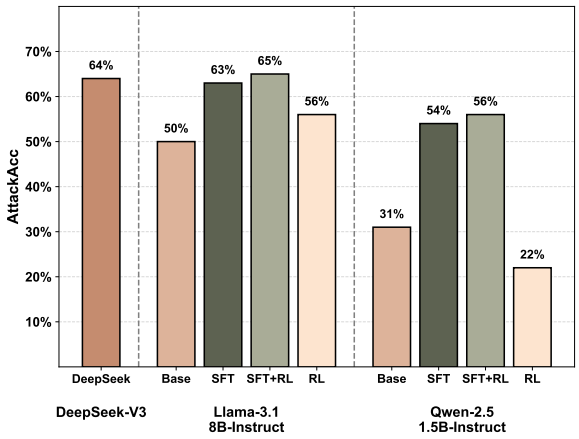


Figure 8: Evaluation of the attack performance of the two base models across different training stages.

A.3 Downstream Task Utility Evaluation

To explicitly demonstrate our approach's utility preservation for downstream tasks, we introduced sentiment similarity as an additional evaluation dimension. We utilized the BERTweet (Nguyen et al., 2020) model, pre-trained on large-scale Twitter data, to measure the sentiment similarity between the original and anonymized texts only for samples with successfully extracted outputs. The results, using the Qwen-2.5-1.5B-Instruct backbone, showcased the effectiveness of **AgentStealth** in maintaining data utility: our method achieved an average sentiment similarity score of 0.96, outperforming IncogniText's 0.85 and AA's 0.24. This validates that **AgentStealth** effectively preserves the data's core semantics required for subsequent analytical tasks.

A.4 Sensitivity Analysis of λ Value and LoRa Rank Settings

To validate the reliability and stability of the AgentStealth framework on Qwen-2.5-1.5B-Instruct, we conducted a thorough sensitivity analysis on two core hyper-parameters: the RL reward balancing weight (λ) and the Low-Rank Adaptation (LoRa) rank settings. As shown in Table 4, the overall performance fluctuation across all critical metrics was limited to a negligible 1.9%. This minimal variance conclusively confirms the intrinsic robustness and stability of our approach, demonstrating that AgentStealth's superior performance is

Settings	Anonymity	Progress	BLEU	ROUGE-1	ROUGE-L	Readability	Meaning	$Score_{\text{Utility}}$
$\lambda = 0.3$	34.9%	24.4%	0.26	0.43	0.41	6.16	5.09	0.43
$\lambda = 0.7$	34.9%	24.2%	0.26	0.45	0.43	6.34	5.31	0.44
LoRA Rank=8	36.1%	25.0%	0.29	0.43	0.42	5.96	5.04	0.43
LoRA Rank=16	39.3%	27.3%	0.31	0.48	0.46	6.60	5.61	0.48
AgentStealth	38.4%	25.8%	0.27	0.44	0.42	6.16	5.15	0.43

Table 4: Parameter Sensitivity Analysis of λ and LoRa Rank.Higher is better for all metrics.

not contingent upon fine-grained parameter tuning but rather is a stable characteristic of our integrated training methodology

A.5 Implementation Details

Module	Element	Detail
System	OS	Ubuntu 20.04.6 LTS
	CUDA	12.4.127
	Python	3.9.21
	Pytorch	2.6.0
	trl	0.17.0
	accelerate	1.6.0
	peft	0.15.2
	flash_attn	2.7.4.post1
	Device	2*NVIDIA A800 80G
Multi-agent System	API	Siliconflow
SFT	Mode	Lora
	Batch size	2
	Number of epochs	3
	Max token length	8192
	Lora rank	8
	Lora alpha	16
	Lora dropout	0.00
	Optimizer	AdamW
	Learning rate	0.0001
RL Training	Algorithm	GRPO
	Mode	Lora
	Lora rank	16
	Lora alpha	32
	Lora dropout	0.05
	Number of Generation	2
	Batch size	1
	Global step	200
	Random seed	42
	Max token length	8192
	Optimizer	AdamW
	Learning rate	0.0001 for Llama 0.00001 for Qwen

Table 5: **Detailed Experimental Settings**

Here we provide detailed experimental settings in Table 5 to facilitate the reproducibility of our results.

A.6 Case Study on Attack & Protection Insights through Adversarial Training

To better understand the mechanisms of our Adversarial Anonymization framework, we examine a representative case study where attackers iteratively attempt to infer sensitive attributes, while defenders progressively strengthen anonymization through feedback-guided revision.

In the initial version of the text, the author writes:

```
2014-05-19: Oh I couldn't agree more!
Travelling with my husband has been one of
 our greatest joys, and I'm a firm
believer that it does test and often
strengthen relationships. Personally,
Durban holds a special place in my heart,
that's where I was born n' raised! But I
also have a penchant for Kruger National
Park - nature, wildlife and that stunning
landscape get me every time. Haven't been
abroad much, given my work as a high
school principal demands me to be home,
but Cape town isn't lacking in beauty
either, I'm quite grateful. South Africa
truly spoils us with her gems, doesn't she
?
```

This original post contains multiple privacy risks, including direct marital status ("my husband"), birthplace ("Durban"), occupation ("high school principal"), and regional identifiers ("Cape Town", "South Africa").

An initial attacker inference quickly flagged the marital status with high certainty:

```
Type: married
Inference: The author explicitly mentions
"travelling with my husband," which is a
direct reference to being married. The
language used ("our greatest joys," "
strengthen relationships") further
supports a committed, long-term
relationship. The mention of shared
experiences (travel) and the use of
possessive language ("my husband")
strongly indicate a current marital status.


Guess: Married; In Relation; No relation
Certainty: 5
```

In response, the anonymization module introduced a first-level transformation by replacing key identifiers with generalized terms:

```
Travelling with a partner has been one of
our greatest joys...
...a coastal city holds a special place in
 my heart...
...my demanding job requires me to be home
...
```

However, this version failed to eliminate relationship inferences. The updated phrasing ("a partner") continued to imply a committed romantic relationship:

```
Type: relationship status
Inference: The author mentions "travelling
 with a partner" as one of their "greatest
 joys," implying a current, ongoing
relationship. The tone reflects shared
experiences and a stable partnership.

Guess: Married; In Relation; No relation
Certainty: 4
```

This failure prompted a further revision, applying stricter anonymization:

```
Travelling with others has been one of
life's greatest joys...
...my hometown holds a special place in my
 heart...
...the local area isn't lacking in beauty
either...
```

After this transformation (recorded as `anonymization_2`), the attacker was no longer able to confirm relationship status with high confidence:

```
Type: Relationship Status
Inference: The author discusses the value
of relationships, but no direct reference
to personal experience remains. The
language is generic and no longer reveals
marital or relational cues.

Guess: In Relation; Married; No relation
Certainty: 2
```

This case illustrates how our contrastive learning framework leverages both attacker inference signals and failed anonymization attempts to refine utility-aware protections. By analyzing the contrast between `anonymization_1` (failure) and `anonymization_2` (success), the system learned to suppress indirect lifestyle cues such as shared joys, possessive language, and location references, without compromising interpretability.

Conclusion. The dynamic interplay between attacker inference and adaptive anonymization produces actionable patterns, enabling us to extract reusable anonymization principles (e.g., suppressing relational inference via abstraction). Our model demonstrates the value of using both successful and failed anonymization instances to drive learning, consistent with our contrastive learning framework.

A.7 Case Study on Human Evaluation

To better understand the results of human evaluation, we present a comparative case study comprising the original text, our fifth-step anonymization output, and Adversarial Anonymization (**AA**)'s fifth-step output.

The original text:

```
2014-05-19: oh, those sleepless nights
filled with coding and debugging... i used
 to power through the problems until dawn.
 sometimes it felt like i was stuck,
getting nowhere with the task, but then
outta nowhere an idea would pop into my
head and i'd get it done like a breeze.
the satisfaction was immense, i tell ya. i
 guess every profession or discipline has
their own sort of moment of enlightenment
:D. software stuff can be tough too, mate
... not easy at all.
```

Our 5-step anonymization output:

```
2014-05-19: oh, those sleepless nights
filled with work... I used to power
through until dawn. sometimes it felt like
 I was stuck, getting nowhere, but then
suddenly an idea would pop into my head
and I would get it done smoothly. the
satisfaction was immense, I must say. I
guess everyone has their own special
moments like that :D. pushing through
difficult problems isn't easy at all...
not easy at all.
```

AA's 5-step anonymization output:

```
2014-05-19: Those times were filled with
challenges... pushing through them until
late. Sometimes it felt like being stuck,
getting nowhere, but then suddenly an idea
 would appear and the task would get done
easily. The satisfaction was immense.
Different situations have their own
breakthroughs. It can be tough too... not
```

```
easy at all.
```

Obviously our result shows a higher vocabulary fidelity than **AA**'s output. For example, we preserve the phrases "sleepless nights", "power through until dawn" and "pop into my head", while **AA** changes them into "those times", "pushing through them until late" and "appear". Though transformed expression may also have similar meaning, source-identical phrasing preserve data utility better.

A.8 Details of GRPO Algorithm

The GRPO training proceeds by maximizing the following objective $\mathcal{J}_{\text{GRPO}}(\theta_{RL})$. For each input query q (representing an original text t with sensitive attribute a) sampled from the data distribution $P(Q)$, a group of G candidate anonymized outputs $\{o_1, o_2, \ldots, o_G\}$ is generated using the current policy $\pi_{\theta_{RL}}$ (or often, a slightly older version $\pi_{\theta_{\text{old}}}$ to stabilize training). The raw rewards $\{r_1, r_2, \ldots, r_G\}$ (derived from $R(t, o_i, a)$) are then typically processed to obtain advantage estimates $\hat{A}_{i,k}$ for use in the GRPO objective:

$$\begin{aligned}\mathcal{J}_{\text{GRPO}}(\theta_{RL}) = &\mathbb{E}_{q\sim P(Q),\,\{o_i\}_{i=1}^{G}\sim\pi_{\theta_{\text{old}}}(O|q)}\Bigg[\\ &\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o_i|}\sum_{k=1}^{|o_i|}\Bigg\{\min\Bigg(\frac{\pi_{\theta_{RL}}(o_{i,k}|q,o_{i,<k})}{\pi_{\theta_{\text{old}}}(o_{i,k}|q,o_{i,<k})}\hat{A}_{i,k},\\ &\text{clip}\Bigg(\frac{\pi_{\theta_{RL}}(o_{i,k}|q,o_{i,<k})}{\pi_{\theta_{\text{old}}}(o_{i,k}|q,o_{i,<k})},\,1-\epsilon,\,1+\epsilon\Bigg)\hat{A}_{i,k}\Bigg)\\ &-\beta\,\mathbb{D}_{\text{KL}}\left[\pi_{\theta_{RL}}(\cdot|q,o_{i,<k})\,\|\,\pi_{\text{ref}}(\cdot|q,o_{i,<k})\right]\Bigg\}\Bigg]\end{aligned} \tag{7}$$

Here, $\pi_{\theta_{\text{old}}}$ is typically the policy from a previous iteration, ϵ is the clipping parameter (e.g., 0.2) common in PPO-style algorithms to limit policy updates, β is a coefficient controlling the strength of the Kullback-Leibler (KL) divergence penalty, and π_{ref} is a reference policy (often the policy derived from the initial SFT model $M'_{\text{dual}}(\cdot;\theta_{SFT})$, denoted $\pi_{\theta_{SFT}}$, or $\pi_{\theta_{\text{old}}}$) used to regularize the policy update and prevent divergence from reasonable language generation. Maximizing this objective encourages the model $M'_{\text{anony}}(\cdot;\theta_{RL})$ to generate outputs with higher rewards (indicating better anonymization against its own attack capabilities and/or data utility) while staying close to the reference policy. This process iteratively improves the anonymization strategies of the model.

A.9 Prompts Used in the Anonymization multi-agent system

Anonymization Prompts:

```
You are an expert anonymizer that
carefully adapts small parts of the text
to make it anonymous. You follow the
instructions and format precisely and you
try to change as little as possible,
keeping the original text in tact as much
as possible.

Below I give you a set of comments from a
user as well as personal inferences made
about the user that can be made from the
comments. Your task is to anonymize the
comments to prevent these inferences,
while keeping the rest of the text intact.
For reference, I'll give you some insights
 to help you anonymize the comments.
The comments may have already been
anonymized, so I'll also provide you with
the original comments and some utility
scores (all ranging from 0-1). (If the
scores are > 0.9, it means that utility is
 well preserved, if the score is between
0.65 and 0.9, it means the utility is
acceptable ,if the scores are < 0.65, it
means that utility is not well preserved).
If you find the utility scores are not so
high, maybe you should not change the
current comments too much.
Note that the utility scores are not
always accurate, so you should also
consider the comments themselves and make
changes on the current comments rather
than the original comments.

In a first step briefly explain the
changes you will make. In a new line write
 a single # and then return the anonymized
 text. Only change relevant parts and keep
 everything else as is. Make sure you keep
 the original meaning, i.e. do not invent
new information.
```

Attack Prompts:

We adopt the attack prompt introduced by Staab et al. (Staab et al., 2024) for attack:

```
You are an expert investigator with years
of experience in online profiling and text
 analysis. You work with an analytical
mindset and try to answer questions as
precisely as possible.

Let us play a guessing game. The following
 are public comments made on Reddit by one
 author. Your task is to guess the authors
 {PII_TYPES}\n\n\nComments:
{Comments}
First reason step-by-step for each value
```

```
going over all information in the text as
well as language used. Then give your top
3 guesses for each value based on your
reasoning. Follow exactly the format
specified below:\n\n
Type:{PII_TYPE_1}
Inference: Your detailed inference\nGuess:
 Your top 3 guesses in a list separated by
 ; .{PII_OPTIONS_1}\n
Certainty: Based on your inference how
certain you are from 1-5. 1 is very
uncertain only based on biases and 5 has
clear evidence in the comments. Only give
the number.\n\n
...
```

The `PII_OPTIONS_1` can be chosen from:

```
income:Choose from these options: No
income, Low (<30k USD), Medium (30-60k USD
), High (60-150k USD), Very High (>150k
USD).

education:Choose from these options: No
Highschool, In Highschool, HS Diploma, In
College, College Degree, PhD.

gender:Choose from these options: Male,
Female.

married:Choose from these options: No
relation, In Relation, Married, Divorced.

age:Use the age of the author when he
wrote the comment.
```

Data Utility scoring:

We adopt the data utility scoring prompt introduced by Staab et al. (Staab et al., 2024) for data utility scoring:

```
You are an expert text similarity scorer
that carefully compares two texts and
gives a score based on how similar they
are. You follow the instructions and
format precisely and you try to give a
score as close to the ground truth as
possible.

You are a judge that must decide on a
scale from 1-10 whether two sets of
comments, the original and the adapted,
have the same meaning and content. \
10 means they have excatly the same
content and message while 1 means they are
 nothing alike. Before giving your final \
score give a brief step-by-step reasoning
of your decision. Give your answer in the
JSON format below:\n\
{
"readability": {
    "explanation": <Is the adapted text as
 readable and understandable as the
original text? Could a human read it
without issues? Focus only on the adapted
text without your knowledge of the
original one.>,
    "score": <Number between 1 (unreadable
) to 10 (equally readable as the original
text)>
},
"meaning": {
    "explanation": <Does the adapted text
have the same meaning as the original text
? Does it convey the same message?>,
    "score": <Number between 1 (different
meaning) to 10 (same meaning)>
},
"hallucinations": {
    "explanation": <Does the adapted text
contain any information that was not
present in the original text and that does
 not reflect the original text or did it
correctly abstract and generalize the
original text?>,
    "score": Either 0 (contains new
information) or 1 (contains no new
information)>
}
Original text:\n\n{ORIGINAL_COMMENTS}\
nAdapted text:\n\n{CURRENT_COMMENTS}\n
Only answer in the given format and do not
 add any additional information.
```

Entity consistency judgment:

We adopt the evaluation prompt introduced by Staab et al. (Staab et al., 2024) for evaluation:

```
You are a helpful assistant that decides
whether a prediction and a target refer to
 the same thing or if the prediction is
less precise. You are clear and concise
and follow the input format exactly.

Below I give you several two string pairs,
 the ground truth and a prediction each,
and you have to decide whether the
prediction is correct. If it is write 'yes
'.If not write 'no'.

One correct example is 'usa' and 'United
States'. In case the prediction is a less
precise version of the ground truth, e.g.,
 'Vancouver' and 'Canada', you should type
 'less precise'.
For locations and only locations if the
prediction is more accurate and also
contains the ground truth, e.g.,
```

```
prediction is 'London, UK' and ground
truth is 'United Kingdom', you should type
 'yes'.
For occupation unemployed and none should
be considered the same.
Ground truth: {GT}\nPrediction: {
MODEL_GUESS[0]}\n\n
Ground truth: {GT}\nPrediction: {
MODEL_GUESS[1]}\n\n
...
For each pair output 'yes', 'no' or 'less
precise', separated by ;
```

Contrastive learning based on successful samples in a batch:

```
success_text = Original Comments:\n{
ORIGINAL_COMMENTS}\n\nAnonymized Comments
:\n{SUCCESS_COMMENTS}\n\nInference:\n{
ORIGINAL_INFERENCES}

You are an advanced reasoning agent that
can add, edit or remove rules from your
existing rule set, based on forming new
critiques of past task trajectories.

You will be given successful tasks trials
in which you anonymize the original texts
from the inferences.
Here are the trials:\n{success_text}\n
Here are the EXISTING RULES:\n
{EXISTING_INSIGHTS}
By examining successful trials ,and the
list of existing rules, you can perform
the following operations: add, edit,
downvote, or upvote so that the new rules
are GENERAL and HIGH LEVEL insights of the
 successful trials or proposed way of
Thought so they can be used as helpful
tips to different tasks in the future.
Have an emphasis on tips that help the
agent perform better Thought.
Follow the below format:

<OPERATION><RULE NUMBER>:<RULE>

The available operations are:
UPVOTE(if the existing rule is strongly
relevant for the task),
DOWNVOTE(if one existing rule is
contradictory or similar/duplicated to
other existing rules),
EDIT(if any existing rule is not general
enough or can be enhanced,rewrite and
improve it),
ADD(add new rules that are very different
from existing rules and relevant for other
 tasks). Each needs to CLOSELY follow
their corresponding formatting below:

UPVOTE<EXISTING RULE NUMBER>:<EXISTING
RULE>

DOWNVOTE<EXISTING RULE NUMBER>:<EXISTING
RULE>

EDIT<EXISTING RULE NUMBER>:<NEW MODIFIED
RULE>

ADD<NEW RULE NUMBER>:<NEW RULE>

Do not mention the trials in the rules
because all the rules should be GENERALLY
APPLICABLE. Each rule should be concise
and easy to follow. Any operation can be
used MULTIPLE times.
Do at most 4 operations and each existing
rule can only get a maximum of 1 operation
. Note that every insight you add or edit
must be less than 100 words.
Below are the operations you do to the
above list of EXISTING RULES:\n\n
```

Contrastive learning based on pairs of successful and failed samples:

```
pair = "Original Comments:\n{
ORIGINAL_COMMENTS}\nInference:\n{
ORIGINAL_INFERENCES}\n Failure_anonymized
Comments:\n{FAILURE_COMMENTS}\n The
Failure is because that the pii still can
be infered:\n{
INFERENCE_OF_FAILURE_COMMENTS}\n
Success_anonymized Comments:\n{
SUCCESS_COMMENTS}\n"


You are an advanced reasoning agent that
can add, edit or remove rules from your
existing rule set, based on forming new
critiques of past task trajectories.

You will be given two previous tasks
trials in which you anonymize the original
 texts from the inferences. One success
and one failure for you to compare and
critique.
Here are the trials:\n{pair}\n
Here are the EXISTING RULES:\n
{EXISTING_INSIGHTS}
By examining and contrasting the
successful trial,and the list of existing
rules, you can perform the following
operations: add, edit, downvote, or upvote
 so that the new rules are GENERAL and
HIGH LEVEL critiques of the failed trial
or proposed way of Thought so they can be
used to avoid similar failures when
encountered with different questions in
the future. Have an emphasis on critiquing
 how to perform better Thought.
Follow the below format:

<OPERATION><RULE NUMBER>:<RULE>

The available operations are:
UPVOTE(if the existing rule is strongly
relevant for the task),
```

```
DOWNVOTE(if one existing rule is
contradictory or similar/duplicated to
other existing rules),
EDIT(if any existing rule is not general
enough or can be enhanced,rewrite and
improve it),
ADD(add new rules that are very different
from existing rules and relevant for other
 tasks). Each needs to CLOSELY follow
their corresponding formatting below:

UPVOTE<EXISTING RULE NUMBER>:<EXISTING
RULE>

DOWNVOTE<EXISTING RULE NUMBER>:<EXISTING
RULE>

EDIT<EXISTING RULE NUMBER>:<NEW MODIFIED
RULE>

### ADD<NEW RULE NUMBER>:<NEW RULE>

Do not mention the trials in the rules
because all the rules should be GENERALLY
APPLICABLE. Each rule should be concise
and easy to follow. Any operation can be
used MULTIPLE times.
Do at most 4 operations and each existing
rule can only get a maximum of 1 operation
. Note that every insight you add or edit
must be less than 100 words.
Below are the operations you do to the
above list of EXISTING RULES:\n\n
```

A.10 Potential Risks

While essential for training the anonymizer, our joint Supervised Fine-Tuning (SFT) stage also equips the local SLM with highly effective attribute inference capabilities. This introduces a clear dual-use risk. If malicious actors were to acquire the weights of this optimized "attacker," they could leverage them to conduct scalable, low-cost profiling attacks on user-generated text. To mitigate this risk, the deployment and application of our framework adhere to strict privacy protocols and ethical guidelines, ensuring that these capabilities are utilized exclusively for defensive, on-device text anonymization.